© 1989 Elsevier Science Publishers B.V. (Biomedical Division)
Present and future of indoor air quality.
C.J. Bieva, Y. Courtois, M. Govaerts, editors.

AN EXAMPLE OF EXTRA-POISSON VARIATION SUGGESTING AN UNDER-SPECIFIED MODEL.

S. James KILPATRICK

Medical College of Virginia, Richmond, Va., U.S.A. 23298-0032

INTRODUCTION

There is now an extensive literature on environmental tobacco smoke (ETS). Several meta-analyses of this literature have combined risk estimates for lung cancer from different studies weighted by the quality and size of the study. Thus, in their meta-analysis, Letzel et al. (1988) give various scenarios which use different figures from Hirayama's study, depending on whether the husband's or wife's age was used to adjust the wife's cumulative mortality. Since Hirayama's study constitutes about 20% of all lung cancer deaths in the ETS literature, it is important to use the correct relative risk from Hirayama in constructing global estimates. Here, it is shown that a proper analysis of Hirayama's study leads to a non-significant association between ETS and the risk of death from lung cancer in non-smoking wives.

Hirayama (1981) reports an age standardised risk ratio of 1.90 for lung cancer among non-smoking women married to heavy smokers and finds a highly significant trend ($P < 0.001$) between the amount smoked by husbands and the lung cancer mortality of their non-smoking wives. He also interprets this association as arising from a causal relationship between husband's smoking and wife's lung cancer:

> These results indicate the possible importance of passive or indirect smoking as one of the causal factors of lung cancer.
> (Hirayama, 1981)

THE STUDY

Dr. Hirayama's study links deaths from all causes occurring in the period 1966 -1981 to a questionnaire given in late 1965 to approximately 250,000 Japanese adults. Almost all those over age 40 in 29 Health Center Districts in 6 Japanese prefectures who were 'generally healthy' (Hirayama, 1978) were interviewed. At that time, each respondent was asked for his or her current smoking status. Although not originally designed as a study of the association between passive smoking and health, Dr. Hirayama subsequently linked the mortality of wives who were classified as non-smokers with their husband's smoking status and, if he was a smoker, with the amount smoked.

Hirayama's smoking classification is based on only one question at the time of the survey as to whether the respondent smoked, and, if so, smoked daily. The respondent was also asked whether he/she smoked occasionally, was an exsmoker, a non-smoker or an 'obscure' smoker. The age at which smoking started was recorded where appropriate. Husband's smoking status in 1965 was treated as an index of wife's ETS exposure. No direct measure of ETS exposure was made. In the following, ETS refers to this surrogate for passive smoking, i.e. the self-reported classification of a husband's smoking in 1965 when his wife was reported to be a

2023381543

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

non-smoker. In Hirayama's 1981 paper there was some ambiguity as to how he had treated the data. The estimates made by Hirayama of the association of ETS and lung cancer lead him to conclude:

> The relation between the husband's smoking and the wife's risk of developing lung cancer showed a similar pattern when analysed by age and occupation of the husband. (Hirayama 1981)

In response to his 1981 paper, a number of methodological questions were raised. In particular, Hirayama's use of the husband's age to adjust the wife's mortality was questioned.

> It is also not clear ... whether you standardised on the age of the wives themselves. Such calculations ... would certainly make the analysis more conclusive. (Harris & Du Mouchel, 1981)

Some years later, Hirayama responded to this criticism by analysing the wife's lung cancer mortality adjusted by the wife's age. In Hirayama (1984) the relative risk of 1.9, reported earlier, for non-smoking wives married to 'heavy' cigarette smokers drops to 1.7 when adjusted by wife's age. He concludes:

> There was a statistically significant increased risk [of lung cancer among non-smoking wives] in relation to the extent of the husband's smoking ... the association was significant when observed by age of husbands (table 1) and also by age of wives (table 2). (Hirayama, 1984 p.179)

THE DATA

Hirayama (1984) gives the lung cancer death rates in the period 1966-1981 for 91,540 self-reported non-smoking wives cross-classified by husband's smoking and by husband's age (Table 1) or wife's age (Table 2). In both tables, age is given in four 10 year age groups. Table 1 gives husband's smoking classified into 5 levels by the amount smoked daily.

TABLE 1

LUNG CANCER DEATH RATES per 1000 (1966-1981)
from Hirayama (1984)

		HUSBAND'S SMOKING				
		Non	Ex	1 – 14/*d*	15 – 19/*d*	20 + /*d*
HUSBAND'S AGE	40–49	0.6	0.8	0.9	1.2	1.5
	50–59	1.3	1.6	2.1	2.0	2.4
	60–69	2.5	4.1	3.9	3.6	4.9
	70–79	6.6	5.7	3.3	9.5	4.4

In order to make direct comparisons between the effect of adjusting by husband's or wife's age, Table 1 is collapsed to Table 1A, using the same grouping for husband's smoking as in Table 2, here called Non, Light and Heavy.

2023381544

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

TABLE 1A

LUNG CANCER DEATH RATES per 1000 (1966-1981) AND SAMPLE SIZE n (000) adapted from Hirayama (1984) Table 1

Husband's	Husband's smoking					
	Non		Light		Heavy	
Age	Rate	n	Rate	n	Rate	n
40-49	0.6	6.2	1.0	15.0	1.5	10.8
50-59	1.3	7.8	2.0	15.6	2.4	9.8
60-69	2.5	7.1	3.9	12.4	4.9	4.7
70-79	6.6	0.8	4.7	1.1	4.4	0.2

TABLE 2

LUNG CANCER DEATH RATES per 1000 (1966-1981) AND SAMPLE SIZE n (000) from Hirayama (1984)

Wife's	Husband's smoking					
	Non		Light		Heavy	
Age	Rate	n	Rate	n	Rate	n
40-49	0.5	7.9	1.2	17.5	1.7	12.6
50-59	1.8	7.6	2.9	15.6	3.5	8.8
60-69	2.6	6.2	3.0	10.4	2.6	3.8
70-79	17.4	0.2	1.5	0.7	8.4	0.2

Note that, in this paper, Table 1A and Table 2 of Hirayama (1984) are both 4 by 3 contingency tables, based on the mortality experience of the same 91,540 women.

MODELLING

The analysis of Table 1A and 2 is usually done using a fixed effects loglinear or logistic model (Breslow & Day 1986). These in turn are viewed as particular examples of a generalized linear model which may be misspecified in three ways

- the linear predictor may be incomplete or incorrect and or
- the wrong link may be used and/or
- the wrong error structure may be assumed.

In epidemiological investigations, the linear predictor is often incomplete in that important covariates are omitted or measured from the wrong origin or in the wrong scale. Such is likely to be the case here in that no allowance is made for diet, for cohort or period effects or for the duration and amount of ETS exposure, either before or after 1965.

As its name implies the loglinear model uses a linear predictor to fit the logarithm of the cumulative mortality rates. Here it is shown that a model allowing for extra-Poisson variation is necessary and that this model leads to a non-significant effect for ETS in Table 2.

The usual regression model relating interval or measurable quantities such as height and weight, assumes a normal distribution of errors about the regression line. By contrast, the loglinear model assumes a Poisson distribution of errors between the observed and fitted

2023381545

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

log(rates). The Poisson distribution, in turn, implies that the mean of the distribution equals the variance or dispersion of the distribution. In other words, the usual model for mortality rates estimates only the mean. Recently, this assumption has been called into question.

> Techniques like regression analysis ... have traditionally focussed attention on modelling and analysis of means or location parameters. Scale parameters have been regarded as nuisance parameters; interest in them has been largely limited to testing the equality of variances so that techniques that assume variance homogeneity can be applied. Recently, however, there has been more interest among statisticians in the structural modelling of variances and in the estimation of dispersion effects. (Anon., the American Statistical Association, 1988)

In the same year in which Hirayama published his only tabulation in which wife's age is used (Table 2 of Hirayama, 1984), Breslow (1984) published a method by which an extra term σ^2 could be added to the loglinear model to allow for excess variation beyond that assumed by a fixed effects model. The model for extra-Poisson variation is shown below to be an extension of the loglinear model.

If d_i deaths are observed among n_i wives, such that $E(d_i) = n_i\lambda_i$ for different subgroups i, then the loglinear model assumes that d_i follows a Poisson distribution with mean $n_i\lambda_i$, where d_i is related to a linear predictor via a logarithmic link or transformation

$E[ln(d_i/n_i)] = x_i\beta.$

In contrast, the extra-Poisson variation model assumes that

$E(d_i) = \mu_i \approx exp[ln(n_i) + x_i\beta]$

where β is a column vector of unknown regression parameters and where

$var(d_i) \approx \mu_i + \sigma^2\mu_i^2.$

The extra-Poisson variance, σ^2, is determined by an iterative re-weighting technique (Breslow 1984). Note that if there is no over-dispersion, σ^2 is estimated as zero.

RESULTS

The usual loglinear model fits Table 1A better than Table 2. Conventionally, a good fit is one for which the residual deviance is less than the degrees of freedom remaining after fitting the linear predictor. The residual deviance after fitting the same linear predictor, age plus husband's smoking level, is 2.0 in Table 1A and 10.7 in Table 2. Since Table 1A has been configured to have the same dimensions as Table 2, both of these deviances may be evaluated against 6, the number of degrees of freedom which remain after fitting age and husband's smoking. In summary then, the deviance for Table 1A is 2.0 with 6 d.f. as compared with a deviance of 10.7 with 6 d.f. for Table 2.

This is paradoxical. Why should the use of the wife's age give a worse fit than the husband's age when modelling the wife's lung cancer mortality? Since this is contrary to our expectation, it suggests that the model is incorrect in one or more of the three categories listed above: incorrect linear predictor, link or error structure.

To allow for the possibility of over-dispersion, Table 1A and Table 2 have been analyzed using the extra-Poisson loglinear model. σ^2 is estimated as zero in Table 1A and as 0.23 in Table 2. An approximate test for the significance of this estimated over-dispersion is to compare the square root of the reduction in deviance after fitting σ^2 with a one-sided Normal distribution. Such a test has an associated probability of 0.015. There is thus some support for the use of this model (and its conclusions) in preference to the fixed effects model. Moreover,

2023381546

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

It is not necessary that σ be significant in order to stay in the model. Often the data are suggestive a priori of excess variation and a prudent analysis will account for it even if not significant. (Mauritsen, 1988)

TABLE 3

EFFECT OF MODEL CHOICE ON P VALUES FOR ETS AS A FACTOR OR A TREND

MODEL	LOGLINEAR	EXTRA-POISSON
FACTOR	.01	.01
TABLE 1A		
TREND	.003	.003
FACTOR	.05	.63
TABLE 2		
TREND	.02	.37

Table 3 shows that this change of model has no effect in Table 1 in which husband's age is used. ETS is statistically significant, both as a factor ($P = 0.01$) and as a trend ($P = .003$) in Table 1A. Nevertheless, the change of model has a marked effect in Table 2 in which the wife's age is used. The ETS factor in Table 2 which is barely significant ($P = 0.05$) under the loglinear model is clearly not significant under the extra-Poisson model ($P = 0.63$). It is inappropriate to test for a trend before the factor has been shown to be significant, in the absence of a prior hypothesis for trend. However, here, following Dr. Hirayama, the trend of the relative risk against the level of ETS is tested. The ETS trend of log (relative risks) with husband's smoking (non, light and heavy) which was just significant ($P = 0.02$) is now, under the model for extra-Poisson variation, not a significant trend ($P = 0.37$).

TABLE 4

EFFECT OF MODEL CHOICE ON 95% CONFIDENCE LEVELS FOR ETS RELATIVE RISKS

MODEL	LOGLINEAR	EXTRA-POISSON
LIGHT	0.98-2.08	0.98-2.08
TABLE 1A		
HEAVY	1.24-2.81	1.24-2.81
LIGHT	0.92-1.97	0.49-2.51
TABLE 2		
HEAVY	1.10-2.48	0.64-3.41

An alternative way of showing the effect of making the model more general is to display the 95% confidence limits for specific levels of ETS exposure. As is seen in Table 4, the model choice has no effect on 95% confidence limits in Table 1A, being 0.98 – 2.08 for the Light vs Non relative risk and 1.24 – 2.81 for the Heavy vs Non relative risk. The choice of model however does affect the 95% confidence limits in Table 2. Here, the Light vs Non relative risk changes from 0.92 – 1.97 to 0.49 – 2.51, neither of which are statistically significant, since both contain the non-effect level of 1.0. The Heavy vs Non relative risk in Table 2 changes from 1.10 – 2.48 to 0.64 – 3.41, so that the interpretation of this association also changes from statistically significant to nonsignificam. Thus, even when the husband reports smoking 20– cigarettes daily, the non-smoking wife shows no significantly increased risk from passive smoking.

2023381547

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

TABLE 5

EFFECT OF MODEL CHOICE ON 95% CONFIDENCE LIMITS FOR ESTIMATES OF ETS TRENDS

MODEL		LOGLINEAR	EXTRA-POISSON
TABLE 1A	TREND	1.11-1.66	1.11-1.66
TABLE 2	TREND	1.05-1.56	0.80-1.87

As shown in Table 5, the 95% confidence limits are consistent with the P values for trend given above. No change is observed in the 95% confidence limits, 1.11 – 1.66, for the trend relative risk in Table 1A. The previously significant trend in Table 2, with 95% confidence limits of 1.05 – 1.56, becomes non-significant with limits of 0.80 – 1.87.

DISCUSSION

Hirayama's analysis assumes that the risk of a lung cancer death is constant within each of the twelve age/exposure sub-groups of Table 2 over the period 1966-1981 and yet:

> Environmental variables are ... difficult to quantify since individual histories vary widely with respect to the onset, duration and intensity of exposure and whether it was continuous or intermittent. (Breslow & Day 1980)

Unfortunately, the danger of using a fixed effects model when unwarranted is that the error term is underestimated.

> Tests of significance and confidence intervals that fail to account for the lack of fit of a given model may be seriously misleading. (Breslow 1987)

Such is the case here. Dr. Hirayama's use of a fixed effects model which gives a poor fit has resulted in his reporting a spuriously significant result. Fitting a more general model confirms that extra-Poisson variation is present in Table 2 (where the wife's age is used) though not in Table 1A (where the husband's age is used instead).

The finding of excess dispersion in this longitudinal record linkage study is likely to be due to the omission of period and cohort effects from the model. Osmond & Gardner (1989) have shown that

> When the assumptions [in the model] are inappropriate, as is usually the case, misleading results will occur.

The period 1966-1981 saw an increasing use of cigarettes in Japan (Kristen 1986) and an increasing mortality from lung cancer in women so that it is unlikely that the association between ETS and lung cancer remained constant over this 16 year period, as is assumed in Hirayama's analysis.

The non-independence of observations within subgroups effectively reduces the sample size, increases the variance and widens the 95% confidence limits for the relative risk of ETS. This is illustrated by the fit of a model allowing extra-Poisson variation, since now the 95% confidence limits include 1.0. In other words, with this model, ETS is no longer significantly associated with lung cancer mortality in non-smoking wives.

2023381548

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

No change is observed, however, after fitting the more general model to Table 1A. ETS is still significantly associated with female lung cancer mortality. This is interpreted as being due to the substitution of husband's age for wife's age. Thus, while the linear predictor used in Table 2 (Wife's Age – Husband's Smoking) demonstrates the absence of period and cohort terms through the need to estimate an extra-Poisson variance component, the linear predictor used in Table 1A (Husband's Age – Husband's Smoking) does not. Rather a model which uses Husband's Age is seen as mimicking a model which includes
(Wife's Age + Husband's Smoking + Wife's Cohort + Period Effects).

Recent research supports this interpretation. Logue & Wing (1986) show that record linkage studies which use rates cumulated over 20 years can produce just the effect reported here, an age/exposure interaction. Under these circumstances, the significance of ETS in Table 1A should *not* be interpreted as indicating a causal relationship but simply that the husband's age, together with husband's smoking status, are proxies for other important but unnamed determinants of an observational study over time.

SUMMARY and CONCLUSIONS

If the wife's mortality from lung cancer is adjusted by the husband's age, the loglinear model gives a good fit and husband's smoking has a significant association with lung cancer mortality in non-smoking wives. It is, however, better to adjust by a person's own age. A consequence of using the wife's age to adjust the wife's mortality from lung cancer is that a loglinear model with extra-Poisson variation is required. With this model, the risk factor, 'husband's smoking', is not statistically significant, indicating no increased risk of lung cancer mortality in non-smoking wives of smoking husbands.

Because of this finding, it is important to use the correct analysis of Hirayama's study in future meta-analyses of published ETS studies in order to get a global estimate of the association of spousal smoking with lung cancer death rates in non-smokers. Such an analysis should incorporate the wife's age when the wife's lung cancer mortality is analyzed.

REFERENCES

- Anon (1988) Amstat News July-August 147, 1
- Breslow NE (1984) J Roy Stat Soc. C 33, 1. 38-43
- Breslow NE (1987) German-American Seminar on Longitudinal Studies and their Analyses. Inst for Social Med
- Breslow NE & Day NE (1980) Statistical Methods in Cancer Research
  Vol I- The Analysis of Case-Control Studies
  I A R C Sc Pub No 32
- Breslow NE & Day NE (1986) Statistical Methods in Cancer Research
  Vol II- The Design and Analysis of Cohort Studies
  I A R C Sc Pub No 82

2023381549

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000

- Harris, JE & Du Mouchel WH (1981) Br Med J 283, 915-916, 3 Oct
- Hirayama T (1978) Nutrition & Cancer 1, 3, 67-81
- Hirayama T (1981) Br Med J 282, 183-185, Jan 17
- Hirayama T (1984) In: Mizell M & Correa P (eds) Lung Cancer: Causes and Prevention Verlag Chemie Int p 175-195 Ch 14
- Kristen MM (1986) Int J Epidemiology 15, 1, 140-141
- Letzel H, Blümner F & Überla K (1988) In: Indoor & Ambient Air Quality, Perry R & Kirk PW (eds) Selper, London pp 293-302
- Logue EE & Wing SJ (1986) Chron Dis 39, 709-717
- Mauritsen R (1988) EGRET Users' Manual Part 3 p 77 S E R C Seattle, Wash
- Osmond C & Gardner MJ (1989) Am J Epidemiology 129, 31-35

2023381550

Source: https://www.industrydocuments.ucsf.edu/docs/pznx0000